Ant Colony based Cloud VM Allocation and Placement Approach for Resource Management in Cloud

Anupama Tiwari
BITS, Bhopal

Pankaj Richhariya
BITS, Bhopal

Satyaranjan Patra, PhD
BITS, Bhopal

ABSTRACT
Resource management in cloud is been an attraction of all the researchers in the past few years. The main reason behind that is the complexity of resource management problem is high. Many virtual machines are created on top of physical machines and this allocation is a NP-Hard problem. In this paper a resource allocation mechanism based on ant colony is proposed.

Keywords
Cloud Service Provider (CSP), virtual machines (VMs), VM monitor (VMM), SLA (Service Level Agreements), physical machine (PM)

1. INTRODUCTION
Computing as a service has seen a phenomenal growth in recent years. The primary motivation for this growth has been the promise of reduced capital and operating expenses, and the ease of dynamically scaling and deploying new services without maintaining a dedicated compute infrastructure. Hence, cloud computing has begun to rapidly transform the way organizations view their IT resources. From a scenario of a single system consisting of single operating system and single application, organizations have been moving into cloud computing, where resources are available in abundance and the user has a wide range to choose from. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with service provider interaction or minimal management effort. Here, the end-users need not to know the details of a specific technology while hosting their application, as the service is completely managed by the Cloud Service Provider (CSP). Users can consume services at a rate that is set by their particular needs. This on-demand service can be provided any time. CSP would take care of all the necessary complex operations on behalf of the user. It would provide the complete system which allocates the required resources for execution of user applications and management of the entire system flow. Figure 1.1, depicts the graphical representation of the architecture of cloud computing [1].

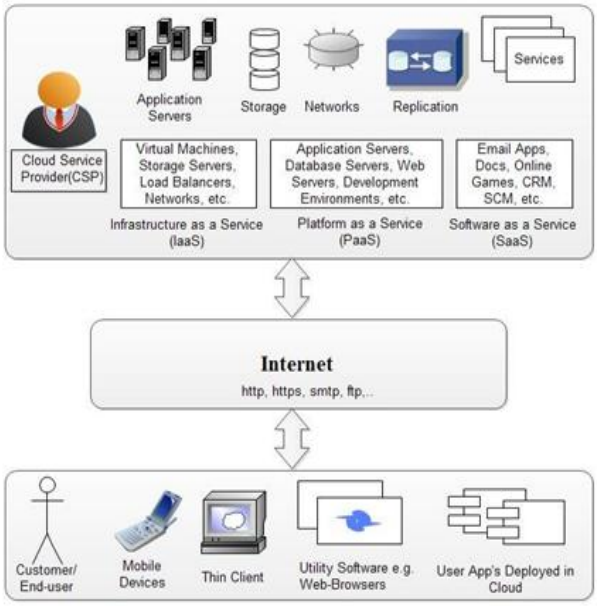


Fig. 1.1 Cloud Computing Architecture

Additionally, this new model has gathered many proponents because of being labeled as a 'Greener Computing Alternative' [2]. Analysts say that pooling of resources and facilities can help cut significant costs for a company. In addition, this also has an extremely positive effect on the environment as an AT&T supported study posits [3]. By 2020, the group estimates, large US companies that use cloud computing can achieve annual energy savings of $12.3 billion and annual carbon reductions equivalent to 200 million barrels of oil.

There has been sources of confusion between grid computing and cloud computing. Clouds and grid have been sharing same visions: reduce computing cost, increase flexibility and reliability. But they differ in the following aspects.

Resource sharing: Grid enhances the share of resources across organizations, whereas cloud provides the resources based on demand of the user. There is no actual sharing due to isolation provided through virtualization.

Virtualization: Grids has capability to virtualize the sum of parts into a singular wide-area resource pool. Virtualization covers both data (databases, flat files) and computing resources. In addition, cloud computing adds virtualization of hardware too.

Security: Cloud Service User has unique access to its single virtualized environment, as virtualization is related to security, where Grid do not deal with end user security.

Coordination: Grids need to perform the coordination of

services workflow and location; whereas in clouds it is not necessary.

Scalability: Grid scalability is mainly enabled by increasing the number of working nodes, whereas cloud resizes the virtualized hardware automatically.

There are a number of advantages for the cloud computing technique that it possesses lower price services, re-provisioning of resources and remote accessibility. Cloud computing lowers the value by avoiding cost by the company in dealing the physical infrastructure from a 3rd party supplier.

In cloud computing the resource allocation possesses associate awfully important role within the performance of the whole system and conjointly the extent of client satisfaction provided by the system. But whereas providing the utmost client satisfaction the service supplier needs to make certain the profits that incur to them conjointly. The resource allocation ought to be economical on each view i.e. on the tip user and therefore the service supplier perspective. Thus on get such a system the new technologies insist that the system ought to be with minimum SLA (Service Level Agreements) violation

SLA: The service level agreement [18] is a part of the terms that is offered by the service provider to give assurance to the end user regarding the level of service that it can provide to the end user. In short, for a customer high QoS suggests few SLA violations

Virtualization is a popular solution that acts as a backbone for provisioning requirements of a cloud-based solution.

Virtualization: Virtualization is the use of hardware and software resources to create the perception that one or more entities exist, although the entities, in actuality are not physically present. Using virtualization, we can make one server appear to be many, a desktop computer appear to be running multiple operating system simultaneously, many network connection appear to exist, or a vast amount of disk space or a vast number of drives to be available. The ability to create virtual machines (VMs) [19] dynamically on demand is a popular solution for managing resources on physical machines.

Virtualization provides a "virtualized" view of resources used to instantiate virtual machines (VMs). A VM monitor (VMM) or hypervisor manages and multiplexes access to the physical resources, maintaining isolation between VMs at all times. As the physical resources are virtualized, several VMs, each of which is self-contained with its own operating system, can execute on a physical machine (PM). The hypervisor [3], which arbitrates access to physical resources, can manipulate the extent of access to a resource (memory allocated or CPU allocated to a VM, etc.).

2. ASPECTS OF RESOURCE MANAGEMENT

A cloud provider's resource management actions toward simultaneously minimizing resource usage and maximizing SLA adherence can be classified as follows:

Load Balancing: There are various resource management policies for balance load in datacenter. The goal of load balancing is to avoid a situation where there is a large discrepancy in resource utilization levels of the PMs. A desired scenario could be to have equal residual resource capacity across PMs (to help increase local resource allocations during increase demands). Virtual machine migrations can be employed to achieve this balance. Load balancing [20] is of two types:

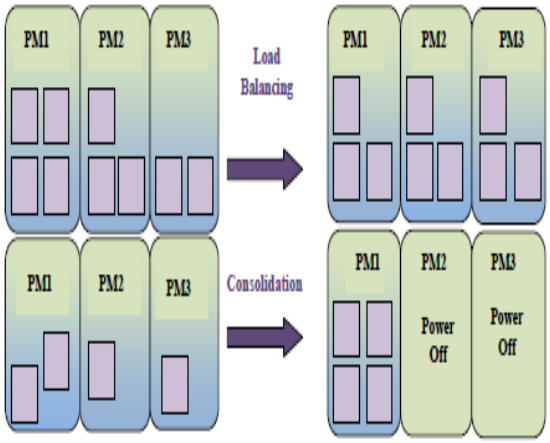
Static Load balancing- In this approach of load balancing, we consider static information of system to choose the least loaded node. It performs better in terms of complexity issue but compromises with the result as decision is made on statically gathered data.

Dynamic Load balancing- In this strategy, current system state plays major role while making decisions. Despite the fact that dynamic load balancing has higher run rime complexity then static one, dynamic has better performance report as it considers current load of system for choosing next datacenter to serve the request.

This will surely provide an optimal choice from available ones for that state of system.

Power Saving: One of the main aspects of Resource management techniques to minimize power consumption at datacenter. To achieve energy efficiency in cloud computing the following methods are useful.

Server Consolidation: The goal of consolidation is to avoid low-resource-usage of host. As shown in Fig, VMs on lightly loaded hosts can be "packed "onto fewer machines to meet resource requirements. The freed-up PMs can either be switched off (to save power) or represent higher-resource availability bins for new VMs.



In the first case, either the goal is to distribute "load" evenly across PMs, or a VM needs more resources and hence is migrated to another PM. With consolidation, machines are migrated to fewer PMs to reduce server sprawl. A single physical virtual server can support 10 or more VMs, allowing numerous applications that normally require dedicated servers to share a single physical server. This facilitates reducing the number of servers in the data center while simultaneously increasing average server utilization from as low as 5-10% up to 60-70%.

Server consolidation is of 2 types.

Static consolidation- The consolidation process can be performed in a single step using the peak load demands of each workload to configure virtual machine capacities, since the virtual machines stay in the same physical servers during their whole lifetime. The utilization of the peak load demand ensures that the virtual machine does not overload. However it can also lead to idleness since the workloads can present variable demand patterns.

Dynamic consolidation- In this reevaluating periodically the workload demand in each virtual machine and performing the required configuration changes and usually results in better consolidation, since it dynamically changes virtual machine capacities according to the current workload demands. However, it may require migrating virtual machines between physical servers in order to:

Pull out physical servers from an overloaded state when the sum of virtual machines capacities mapped to a physical server becomes higher than its capacity, or

To turn off a physical server when the virtual machines mapped to it can be moved to other physical servers. Dynamic consolidation involves VM migration from one host to another host.

3. MIGRATION POLICY AND HEURISTICS

In cloud computing environment, user processes are executed on a virtual machine. Virtual machine migration enables load balancing, power saving, and improving resource usage. In case of using a public cloud system and a private cloud system, virtual machines are migrated into a remote network through a wide area network. Migration performance, including performance of processes on a migrating virtual machine, is one of important issues in the current cloud computing system, which is composed of public and private clouds. In a cloud computing environment, Virtual machine migration [22] is a function that a running virtual machine moves from a physical computer to another computer. It is implemented in several virtual machine systems, such as VMware, Xen, KVM, and Open VZ.

Virtual machines are migrated with two kinds of methods.

Non-live migration: In this a virtual machine stops their processes during migrations and the virtual machine's state is transferred from the source computer to the destination computer. After transferring, the virtual machine restarts in the destination computer. Ex. Suspend-and-copy technique.

Live migration: In a live migration process, a virtual machine begins transferring its state without stopping. When transferring has finished, the virtual machine's state has changed, but their performance may decrease. The difference is additionally transferred. This additional transmission is repeated to decrease the difference. When the difference becomes enough small, the virtual machine stops at the source computer, and transfers the difference. After transmission, the virtual machine restarts at the destination computer. A virtual machine can migrate without stopping its service long time Ex. Pre-copy and Post copy techniques. Live migration techniques, aim to minimize downtime.

The pre-copy approach [23] transfers pages iteratively to the target machine without suspending the VM (and hence is live). Once "sufficient" pages are transferred, the VM is suspended at the source and remaining state transferred to the target machine.

In Post-copy approach the pages are transferred by copying minimal state to start the VM and using demand-paging over the network to fetch the remaining state.

For each of the goals — consolidation, and load balancing, there are three VM migration heuristic.

When to Migrate

There are many situations when migration of VMs becomes necessary to maintain the overall efficiency of the data center. These situations are triggered.

Due to Hot Spot— hot spot is the overloaded condition of a PM. It can also be defined as the state when performance of a system falls below the minimum acceptance level. Hot spot can be detected by analyzing the trends in resource utilizations of the VM. There are some heuristic for VM to migrate.

Fixed Utilization Threshold

Single threshold (ST): ST is based on the idea of setting an upper utilization threshold for hosts and placing VMs while keeping the total utilization of the CPU below this threshold. The aim is to preserve free resources to prevent SLA violation due to consolidation in cases when the resource demand by VMs increases.

Double threshold (DT): DT is based on the idea of setting upper and lower utilization thresholds for hosts and keeping the total utilization threshold of the CPU by all the VMs between these thresholds. If the CPU utilization falls below the lower threshold, all the VMs have to be migrated from this host and the host has to be switched off in order to eliminate the idle power consumption. If the utilization threshold exceeds the upper threshold, some VMs have to be migrated from the host to reduce the utilization in order to prevent potential SLA violation.

4. PROPOSED APPROACH

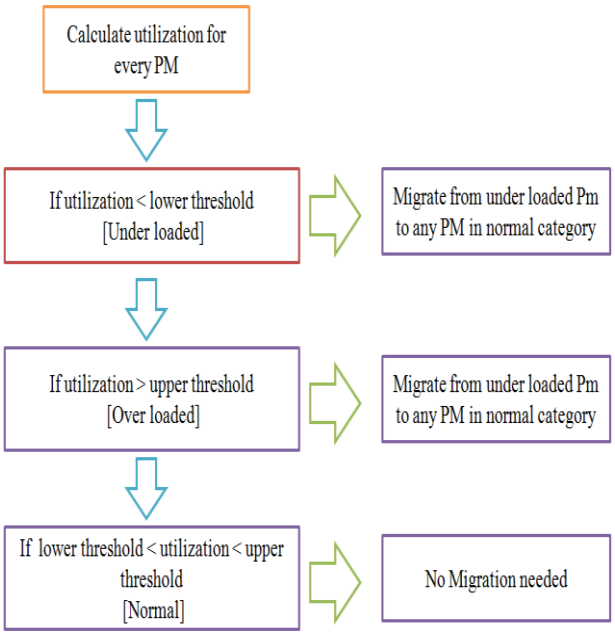
Consider each physical machine is represented by a node in graph and each edge defines VM migration from one physical machine (PM) to another. The generated graph will be directed and completely connected having positive edge weights.

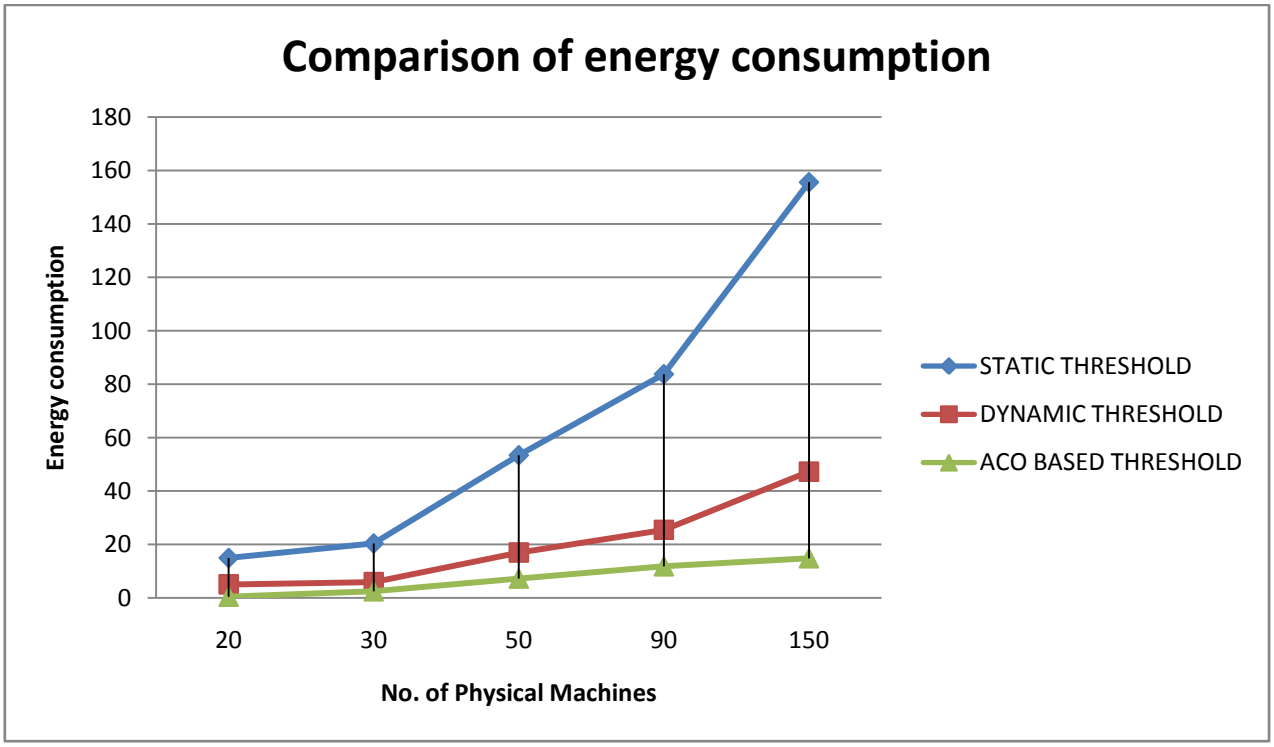
Dynamic load balancing
{
Set upper threshold (UT) to 75% and lower threshold (LT) to 25 % as per standard.
Take lower window (LW) = 70 and upper window (UW) = 90.
Take delta = 2
Calculate average load of cloud datancenter (AvgLoad)
If (AvgLoad > UT)
{
  If (AvgLoad + delta < UW)
    UT = AvgLoad + delta
}
Else if (AvgLoad < UT)
{
  If (AvgLoad + delta > LW)
    UT = AvgLoad + delta
}
}

In AS ants concurrently build the solution for the Cloud VM. Initially ants are put on randomly chosen nodes which represent PM. In each iteration construction step, ant k applies probabilistic action choice rule, called random proportional rule, to decide to which PM given VM should be migrated. Pij is the probability of migrating VM to j which is currently at i is:

$$p_{ij}^{k} = \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l \epsilon N_i^k}[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}} \qquad if\ j \epsilon N_i^k$$

Where ηij = 1/dij is a heuristic. α and β are two parameters which determine the relative influence of the pheromone trail and the heuristic information, and where N_i^k is the nodes which are available.





5. REFERENCES

[1] Komal Singh Patel and A. K. Sarje, "VM Provisioning Method to Improve the Profit and SLA Violation of Cloud Service Providers," IEEE International Conference, Cloud Computing in Emerging Markets (CCEM) 11-12 Oct. 2012.

[2]K. S. Patel and A.K. Sarje, "VM Provisioning Policies to Improve the Profit of Cloud Infrastructure Service Providers," ICCCNT-12, July.2012.

[3]Gundeep Singh Bindra, Prashant Kumar Singh,Seema Khanna,Krishen Kant Kandwal, "Cloud Security : Analysis and Risk Management of VM Images," Proceeding of IEEE International Conference on Information and Automation Shenyang, China, June 2012.

[4] Eeraj Jan Qaisar, "Introduction to Cloud Computing for Developers," In IEEE ©2012.

[5] A. Beloglazov, R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers," Concurrency and Computation: Practice and Experience (CCPE), Wiley Press, New York, USA, Sep. 2012, pp. 1397–1420, doi: 10.1002/cpe.1867.

[6] Zhibo Cao and ShoubinDong , "Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing," 13th International Conference on Parallel and Distributed Computing, Applications and Technologies 2012.

[7] YonggenGu, Wei Zhang, YonggenGu, Jie Tao, "A Study of SLA Violation Compensation Mechanism in Complex Cloud Computing Environment," In IEEE © 2012.

[8] C. Belady, "In the data center, power and cooling costs more than the equipment it supports," 2007. URL http://www.electronicscooling.com/articles/2007/feb/a3/.

[9]http://en.wikipedia.org.

[10]http://www.sciencedirect.com/science/article/pii/S1877705811054117.

[11]David Aikema, AndreyMirtchovski, Cameron Kiddle, and Rob Simmonds "Green Cloud VM Migration: Power Use Analysis" in IEEE 2012.

[12] Saurabh Kumar Garg, Adel NadjaranToosi, Srinivasa K. Gopalaiyengar, RajkumarBuyya, "SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter," Journal of Network and Computer Applications 1 August 2014 .

[13] Rafid Sagban, Ku Ruhana Ku Mahamud, Muhamad Shahbani Abu Bakar "Reactive Memory Model for Ant Colony Optimization and Its Application to TSP" in 2014 IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014, Penang, Malaysia.

[14] M. Veluscek, T. Kalganova, P. Broomhead "Improving Ant Colony Optimization Performance through Prediction of Best Termination Condition" in IEEE 2015.

[15] Fahimeh Farahnakian, Adnan Ashraf, TapioPahikkala,PasiLiljeberg, JuhaPlosila, Ivan Porres, and HannuTenhunen "Using Ant Colony System to ConsolidateVMs for Green Cloud Computing" in IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 8, NO. 2, MARCH/APRIL 2015.

[16] K. Mills, J. Filliben, and C. Dabrowski, "Comparing vm-placement algorithms for on-demand clouds," in Proc. IEEE 3rd Int. Conf. Cloud Comput. Tech. Sci., 2011, pp. 91–98.

[17] H. Xu and B. Li, "Anchor: A versatile and efficient framework for resource management in the cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1066–1076, Jun. 2013.

[18] S. Di and C.-L. Wang, "Dynamic optimization of multi-attribute resource allocation in self-organizing clouds," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 3, pp. 464–478, Mar. 2013.

[19] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in Proc. Conf. Power Aware Comp. Syst., 2008, pp. 10–10.

[20] B. Speitkamp and M. Bichler, "A mathematical programming approach for server consolidation problems in virtualized data centers," IEEE Trans. Serv. Comput., vol. 3, no. 4, pp. 266–278, Oct. 2010.

[21] Sujesha Sudevalayam and Purushottam Kulkarni, "Affinity-aware modelling of CPU usage with communicating virtual machines" in Journal of Systems and Softwares 86 (2013) 2627-2638.

[22] Pablo Graubner, Matthias Schmidt and Bernd Freisleben, "Energy-Efficient Virtual Machine Consolidation" in IEEE Computer Society, IT Pro, March/April 2013.

[23] Tiago C. Ferreto, Marco A. S. Netto, Rodrigo N. Calheiros and Cesar A.F. De Rose, "Server Consolidation with migration control for virtualized data centers" in Future Generation Computer Systems 27 (2011) 1027-1034.